Apr 12th, 2:30 PM - 3:50 PM

Data Mining And Machine Learning Applied To A Software Development Social Collaboration Network

William Lyon
lyonwj@gmail.com

Data Mining and Machine Learning Applied To A Software Development Social Collaboration Network

William Lyon

March 1, 2014

1 Overview

GitHub[1] is a web-based hosting service for software development projects making use of the Git version control system. GitHub provides a framework that allows software developers to easily collaborate on projects and share these projects, releasing them to the world under various open source software licenses. This can be thought of a collaboration network. However, GitHub provides much more than simply supporting the distributed Git version control system for software development. In addition, GitHub has implemented a social network, allowing rich user interactions that extend beyond the typical collaboration network. This project explores how machine learning can be applied to this network to gain insights about its users, projects under development, and the state of software development at large.

2 Methodology

Much of this project focuses on the data modeling requirements necessary to model and store GitHub user activity. Specifically, a graph database (a relatively new type of NoSQL database) is used to store the data describing user interactions. Due to the large volume of data being analyzed[2], a distributed graph database system is used, allowing for the efficient store, retrieval and traversal of the GitHub network. Once the data has been appropriately modeled in the graph database cluster, machine learning algorithms can be applied to gain insight into the network. This presentation will focus on the implementation of a collaborative filtering recommender system[3] which is capable of identifying similar users in the network. Collaborative filtering is a machine learning technique that uses a similarity calculation to make inferences about a dataset by comparing to observations with the highest similarity value. A similarity measure technique described in [4] is used during implementation.

3 Impact

Open source software has become important in fields beyond computer science. Many fields of academic research rely on open source software for analysis tools. Many significant software projects are built on top of open source software. The emergence of GitHub as a major facilitator for collaboration in the open source community, combined with the openness and public accessibility of GitHub data makes it an ideal subject for understanding what is happening in the open source software community. Being able to harness and analyze this data is key to understanding how the open source software development community is developing.

References

[1] URL: `http://github.com`.

[2] URL: `http://developers.github.com/v3/`.

[3] Yehuda Koren and Robert Bell. "Advances in Collaborative Filtering". In: *Recommender Systems handbook* (2011), pp. 146–186.

[4] Francoise Fessant Laurent Candillier Frank Meyer. “Designing Specific Weighted Similarity Measures to Improve Collaborative Filtering Systems”. In: *Lecture Notes in Artificial Intelligence: Advances in Data Mining. 8th Industrial Conference, ICDM 2008.* 5077 (2008), pp. 242–255.